Conception d'un modèle d'indexation de documents structurés multimédias : cas des relations textes - images fixes

Bernadette Monnet Badjo et Marc Berthier

Introduction

Problématique

L'indexation est à l'intersection de deux problématiques : l'existence d'un volume d'information et le besoin d'accès à cette information. L'indexation automatique naît de l'explosion documentaire que connaît le monde ces dernières décennies. Cette explosion est due à l'émergence de nouveaux supports d'information. Ces supports électroniques aux capacités impressionnantes de stockage de différents médias modifient la recherche documentaire en recherche d'information. Dans un ca,s il s'agit de retrouver des documents d'information. Dans l'autre, se pose l'accès direct à l'information des gros volumes de documents intégraux sur ces supports électroniques.

Cette contribution aborde la question au niveau d'un type spécifique de corpus : le document technique (DT) de maintenance. La spécificité de cette indexation réside en outre dans le fait de prendre en compte les ''images''. Sous ce terme, nous regroupons photos, schémas, dessins, organigrammes, graphiques, tableaux, légendes. Les paratextes au sens de G. Genette[1],' c'est-à-dire des éléments textuels satellites du discours principal, sont même inclus.

L'importance de l'enjeu de l'indexation dans le contexte des documents électroniques, comme l'intérêt de l'utilisation de l'image, ne sont plus à démontrer. Trouver les clés pour un accès rapide à une information pertinente au sein de réseaux multiformes et innombrables ou simplement au sein d'un gros document est le but essentiel de l'indexation automatique. Nous ne nous attarderons donc pas ici à définir l'image dont les contours restent difficiles à cerner, ni l'indexation automatique et à les justifier. Nous donnons brièvement par contre les caractéristiques spécifiques des documents structurés multimédias et du DT qui justifient le choix de notre modèle d'indexation. Nous présentons dans une première partie notre modèle d'indexation, avant d'aborder en second lieu les limites de ce modèle et de terminer par l'étude des perspectives d'application à d'autres types de corpus.

L'objet d'étude

Le document structuré multimédia

Les caractéristiques des documents structurés multimédias sont contenus dans les définitions de Balpe (1996) et de Cotte (2000). Pour le premier auteur, « Sous le terme de document structuré, on entend les différentes entités logiques entrant dans l'élaboration d'un document, les règles d'organisation de ces entités entre elles, et les outils de manipulation de ces entités ». Quant au second chercheur, «Le document balisé étant

[1] G. Genette, cité dans http://procyon.inrp.fr/Acces/Innova/Observatoire/Missions/Missions/Web.htm

destiné à une première lecture machinique avant d'être livré à la lecture humaine doit comprendre des signaux redondants destinés à la machine. C'est la fonction d'une balise titre (<TITLE>…<TITLE> dans le langage HTML) par exemple ».

Le support du document électronique et sa lecture par une machine exigent, selon ces définitions, une forme spécifique de ce document dans laquelle la mise en évidence de ses éléments de structure est essentielle.

Le document technique

Sous ce vocable se trouve représentée une gamme variée d'objets de la réalité : texte de loi, rapport financier d'une entreprise, rapport épidémiologique, mode d'emploi d'un appareil, etc. Le document technique auquel nous faisons référence, ici, se rapporte à des ''dispositifs'', i.e. des objets artificiels résultant d'un processus technologique. La documentation technique est un corpus documentaire accompagnant ce dispositif , «un élément des systèmes industriels au même titre que leurs autres composants. En effet leur complexité est telle qu'ils sont accompagnés, durant leur cycle de vie, de documents qui les décrivent, sorte de miroirs sans lesquels, ils ne peuvent ni être fabriqués, ni utilisés » (Froissart 2001).

Suivant le niveau de complexité du dispositif, cette documentation peut être peu importante ou atteindre des volumes considérables. C'est le cas des domaines de l'armement, du nucléaire civil. Le document technique associe aux textes une forte proportion d'images et des paratextes. Il se caractérise par une structuration logique forte, chaque partie formant un tout. Son langage minimise l'argumentation et l'implicite et se caractérise par l'emploi prioritaire de formes verbales actives et de nominalisation. De même, ce langage intègre des formes picturales, symboliques et des langages de spécialités.

L'objectif du document technique est la satisfaction d'un besoin d'information précise pour une action immédiate sur le dispositif. Il ne s'agit donc pas d'indiquer des pages où se trouveraient des informations éventuelles comme pour une recherche d'information de type cognitif, mais de localiser exactement cette information. La conséquence en est que l'indexation doit pouvoir donner les clés d'accès à des unités d'information avec une granularité plus grande.

Notre corpus

Notre corpus se compose de deux manuels de maintenance de dispositifs matériels, imprimantes de la société Bull et centrifugeurs et incubateurs de la société Jouan, et d'un manuel d'utilisation d'un logiciel Sytos d'une entreprise américaine Sytran Corporation. Nous avons également travaillé sur un quatrième document, de vulgarisation en mécanique automobile, pris sur le Web[2]. Dans ce document, il s'agissait de se pencher essentiellement sur les ''images'' pour en dégager leurs caractéristiques spécifiques afin d'élaborer une typologie des éléments non textuels dans le document technique (DT)- non abordé ici- de manière à les intégrer comme entités d'indexation pour la RI.

[2] http://motorlegend.com/moteur2temps

Le but du modèle

Dans le contexte des réseaux numériques, l'indexation automatique cherche à résoudre des problèmes qui sont ceux des temps de traitement de l'information et par la suite d'accès à cette information pour minimiser les coûts. Elle cherche à appliquer à tout document un ensemble de règles fixes permettant d'obtenir les mêmes termes sur un même corpus quel que soit l'indexeur, et enfin à résoudre le problème de surcharge cognitive par la proposition d'informations les plus pertinentes possibles.

Concernant la question du temps, l'indexation automatique doit réduire la durée de traitement dans le repérage des clés d'accès au document électronique par la capacité rapide de calcul de l'ordinateur. Quant à la question de la variation des termes représentatifs d'un même objet, il s'agit plus des multiples synonymes que du nombre de termes différents pouvant être utilisés pour désigner un même objet. Un outil d'analyse efficace appliquant les mêmes règles précises sur chaque document permet d'harmoniser les différentes désignations, minimisant ainsi la subjectivité. Pour le problème de surcharge cognitive, l'identification de termes précis représentatifs d'une réalité constitue une première étape vers la résolution du problème. La mise en évidence de la structure du document qui détermine le segment textuel ou ''paratextuel'' liée à une information prend en compte et traite la question de surcharge cognitive.

Le but à atteindre par le modèle d'indexation proposé pour les DT est la pertinence dans les résultats de la recherche d'information. Cette « pertinence » est évaluée classiquement à l'aide de deux critères, le « taux de rappel » et la « précision ». Dans cette optique, nous proposons une méthode d'indexation basée sur la structure du document et l'analyse linguistique du discours de ces documents.

Proposition d'une méthode d'indexation structurelle et linguistique

Les étapes de la méthode

La méthode s'organise en quatre étapes successives :

-la représentation du document sur la base de la structure d'organisation logique pour localiser par certains indices les différentes catégories d'unités d'information ;
-l'analyse linguistique dans le but d'extraire des candidats descripteurs ;
-l'établissement d'une matrice, en vue d'une analyse statistique sommaire des occurrences des candidats descripteurs ;
-'exploitation de la matrice, pour la construction d'un index en vue de la recherche ultérieure d'information (RI).

La représentation du corpus sur la base de la structure d'organisation logique des documents

Les raisons

Dans la littérature concernant les documents multimédias, la notion de structure est centrale, imposée par la nature même de ces documents dits structurés. Le document technique est par nature, très fortement organisé. Il peut ainsi être appréhendé dans son analyse à partir de sa propre structure.

La notion de structure est elle-même multiple. On parle de structure physique, logique, temporelle, etc. Notre modèle se base sur la structure logique, i.e. l'organisation en parties et sous parties de l'information (par l'auteur du document). Elle est intimement liée à la structure physique qui est la mise en forme matérielle de cette organisation logique. L'objectif de trouver l'information la plus précise possible fonde à elle seule, l'utilisation de l'emboîtement des parties du discours d'un document.

Comment détecter ‘’automatiquement’’ la structure du document ?

Il s'agit de traduire la structure de chaque document à indexer en langage SGML, (Standard General Markup Langage) tout en y incluant les informations sur les images. L'organisation logique, écrite en SGML décrit de la façon la plus neutre possible (sans faire référence à un usage particulier), l'organisation hiérarchique d'un document quelconque. Aucune information de niveau “ méta ” n'apparaît ; seul le contenu du document est représenté. Des éditeurs de documents structurés existent dont GRIF utilisé par l'équipe CRISTAL. GRIF, à partir d'une DTD (Définition type de Document) représente la structure abstraite logique d'un document donné.

C'est lors de cette première étape que des documents d'origines diverses, donc d'organisations logiques diverses et de formats informatiques différents devront être ramenés à un format et une organisation communs. Notre schéma de structure sera le plus simple possible, se limitant à représenter l'emboîtement des parties du document et à distinguer les textes jouant un rôle de titre des autres parties textuelles. Nous utiliserons donc le format d'édition des documents électroniques de façon très élémentaire dans la délimitation de nos différentes unités d'information.

Extension aux documents multimédias

L'édition de documents structurés représente sous forme hiérarchique (arborescente) la structure logique d'un document ou d'un corpus de documents. Le document ou le corpus de documents se décomposent ainsi en parties, ces parties à leur tour sont décomposées en sous-parties, et ainsi de suite. Si le document ou le corpus est strictement textuel, les unités “ atomiques ” , c'est-à-dire considérées comme élémentaires, peuvent être des paragraphes, ou des groupes de paragraphes (avec d'ailleurs des variations possibles dans la définition du paragraphe). Si le document comporte des éléments non textuels, chacun de ceux-ci peut alors être introduit dans la représentation hiérarchique du document, soit comme entité intégrée au paragraphe, soit comme entité élémentaire d'un autre type.

On aurait de plus le moyen de distinguer aisément plusieurs types pour les éléments non textuels de nature différente : photographie, schéma, tableau etc. Nous optons pour faire des éléments non textuels i.e. les images des unités d'information équivalentes aux unités textuelles.

L'analyse linguistique

Les raisons

Pour l'équipe CRISTAL, les textes à analyser sont une manifestation de la langue écrite ou orale, à ce titre il faut privilégier l'analyse linguistique par rapport à l'analyse statistique. L'indexation automatique doit se baser sur un modèle linguistique, parce qu'un descripteur n'est pas une simple unité lexicale de surface, il est lié à un contexte donné. La langue,

bien que soumise à des règles, est par essence ambiguë. Les modèles linguistiques du TAL réorganisent la grammaire usuelle pour construire des outils informatiques opérationnels en palliant l'ambiguïté des textes. Une couverture large de la langue par le modèle conduit à une grande complexité de celui-ci.

Comment effectuer cette analyse ?

L'analyse linguistique comporte deux phases essentielles de traitement : l'analyse morphologique et l'analyse syntaxique. *Dans l'analyse morphologique,* on retient que l*e rôle de l'analyseur morphologique (sous-entendu flexionnel, par opposition à dérivationnel) est d'associer à chaque forme rencontrée en surface, une ou plusieurs interprétations constituées de l'entrée lexicale canonique, d'une catégorie syntaxique, et de diverses variables (syntaxiques, flexionnelles et lexicales). On obtient* un couple (base, flexion), les bases se trouvent dans un dictionnaire, où sont définis leurs comportements flexionnels, c'est-*à-dire un ensemble de règles d'utilisation, sous-jacentes de cette forme, attestés dans la langue. Ces comportements sont saisis à travers des modèles linguistiques. Comme l'indique Lallich-Boidin, Henneron et palmerti (1990 : 19) :* «L'analyseur morphologique combine deux fonctions : une fonction classificatoire, de nature lexicale, qui consiste à attribuer à chaque forme du texte un*e* catégorie syntaxique et une fonction calculatoire ou flexionnelle qui consiste à partir d'une forme donnée, à calculer sa base pour accéder au lexique ».

Exemple : analyse morphologique de la forme *interne* dans la phrase suivante

On dit qu'un moteur alternatif à combustion interne fonctionne selon ...

```
&     F  interne   GRN SNG NAN
      v  interner  PE2 IPF PST ORD
      v  interner  P13 ISU PST ORD
```

La forme *interne* relève de 2 catégories : celle des Nominaux noté F et des verbes noté v. En effet, dans le modèle linguistique utilisé, la classe des nominaux regroupe les noms, les adjectifs, les participes passés et présents considérés comme adjectifs.

Une variable syntaxique au niveau de la catégorie F, précise que *interne* peut être nom ou adjectif : valeur NAN. Une variable flexionnelle donne le genre : Genre neutre (GRN), et une autre variable donne le nombre: singulier (SNG).Une variable syntaxique au niveau de la catégorie v distingue le verbe « ordinaire » *interner* d'un auxiliaire par la valeur ORD. Une variable flexionnelle porte sur la personne du verbe (2è, soit PE2 ou 1° ou 3°, soit P13), son temps (présent soit PST), son mode (impératif soit IPF ou indicatif ou subjonctif ISU).

Le premier objectif de l'utilisation de l'analyseur morphologique est de reconnaître une même unité lexicale sous différentes flexions. Mais l'interprétation morphologique est fortement ambiguë vue qu'elle se réalise sans considération de contexte.

En entrée de l'analyse syntaxique, ne peuvent être utilisées toutes les interprétations, il faut effectuer un filtrage afin de proposer la séquence adéquate : une étape de désambiguïsation statistique de type markovien (ordre 2) peut permettre de choisir l'interprétation comme nom (indication &).

Le principe de cette analyse statistique est très brièvement résumé dans cette citation :

> « Soit un système qui peut prendre un ensemble fini d'états et qui change d'état à chaque instant t. Si la probabilité d'apparition d'un état à un instant donné est fonction d'un nombre constant d'états précédents du système celui-ci peut être qualifié de markovien. On s'intéressera plus particulièrement aux systèmes où la probabilité d'apparition d'un état dépend des n états précédents. On dira qu'ils relèvent d'un modèle markovien d'ordre n (p. 47).

L'application de ce principe à l'analyse linguistique considère les catégories syntaxiques comme les états possibles d'un système et leur succession dans la phrase comme une suite d'états pris par le système. Cela signifie que le système doit déterminer des séquences dans lesquelles certaines catégories apparaissent dans un ordre constant. Cette analyse statistique permet de réduire dans une proportion importante l'ambiguïté.

L'analyse morphologique va ainsi permettre de compter plus tard les mots en faisant abstraction des flexions en postulant que les nombres d'occurrences seront plus significatifs. Le texte initial est ainsi réécrit en un autre langage pour la machine. Elle va être suivie et complétée par l'analyse syntaxique proprement dite, opérant sur ses résultats pour à terme permettre l'extraction des syntagmes en vue de la RI ultérieure. L'analyse syntaxique est précédée de la segmentation des phrases en propositions pour :

> « Limiter la longueur de la chaîne à analyser, et limiter aussi la grammaire de la syntaxe aux seuls phénomènes internes à la proposition. La segmentation s'effectue sans analyse en profondeur,... à partir de l'observation des différentes catégories des formes de surface » (p. 59).

L'analyse syntaxique

Elle consiste à analyser de manière fine la liste des séquences de couples (entrée lexicale, catégorie) correspondant à une proposition, conformément à une grammaire formelle. «Le résultat attendu de l'analyse syntaxique est la (ou les) structure(s) syntaxique(s) des propositions, structures attestées par la langue » (p. 62). Pour notre application à l'indexation, cette technique est trop lourde et produit des résultats trop détaillés induisant de nouvelles difficultés. Nous lui avons substituée une méthode d'analyse dite « robuste » dont nous détaillerons le fonctionnement dans la partie 2.2.

Dans cette étude, nous travaillons spécifiquement sur le document technique. Une caractéristique de ce document est un type de langage dans lequel les règles classiques de construction des phrases sont fréquemment enfreintes. C'est le cas d'énumérations ou listes, de consignes. Véronika Lux propose de faire précéder l'analyse syntaxique d'une étape, supplémentaire celle de transformation des différentes formes de listes qu'elle a identifiées, à un seul format dans un schéma XML. Elle utilise pour cela le logiciel CIRCUS de transformation de structure[3]

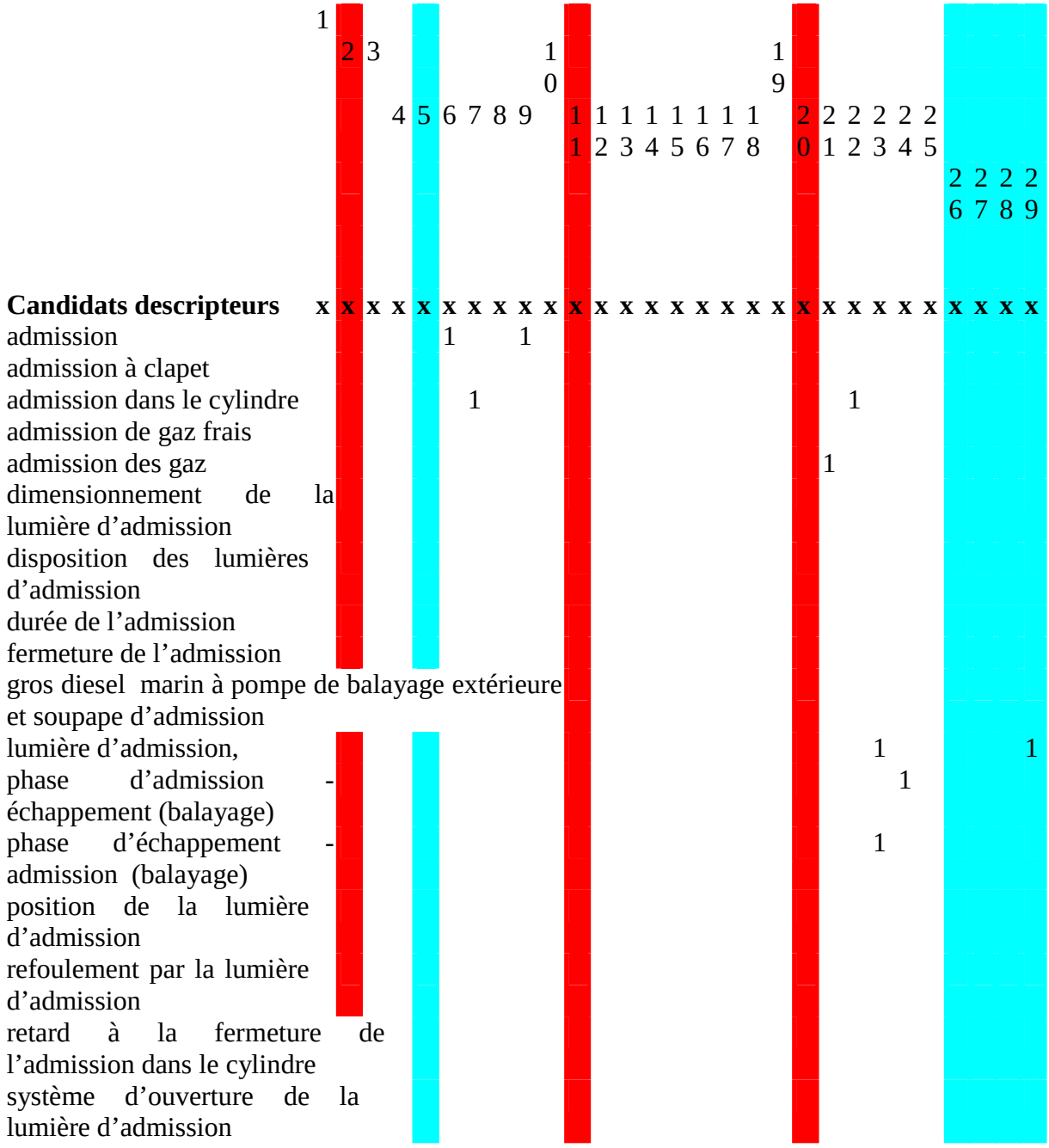
La construction de la matrice des occurrences de candidats descripteurs

À l'issue de l'analyse linguistique (ou de l'analyse « robuste »), les candidats descripteurs sont identifiés. La question est alors de compter les occurrences de ces candidats descripteurs dans les différentes parties du corpus. Ces comptages peuvent être effectués

[3] Site de CIRCUS : http://www.xrce.com.

de différentes manières. Nous adoptons une matrice dont chaque case dénombre les occurrences d'un candidat descripteur dans une partie du corpus.

Les colonnes de la matrice reproduisent les parties du corpus. Les colonnes de niveau supérieur constituent les blocs d'information contenant des groupes d'unités d'information (UI). Les colonnes inférieures, sont formées des UI, les plus élémentaires définis dans la structure hiérarchique du document. Les lignes correspondent aux candidats descripteurs formés de syntagmes simples et complexes.


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Candidats descripteurs x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
admission 1 1
admission à clapet
admission dans le cylindre 1 1
admission de gaz frais
admission des gaz 1
dimensionnement de la lumière d'admission
disposition des lumières d'admission
durée de l'admission
fermeture de l'admission
gros diesel marin à pompe de balayage extérieure et soupape d'admission
lumière d'admission, 1 1
phase d'admission - échappement (balayage) 1
phase d'échappement - admission (balayage) 1
position de la lumière d'admission
refoulement par la lumière d'admission
retard à la fermeture de l'admission dans le cylindre
système d'ouverture de la lumière d'admission
Extrait de la matrice pour le document Motorlegend

Cette matrice est une nouvelle représentation du document initial après le rejet des ''mots dits vides'' et des formes verbales. À cette troisième étape de la méthode, les candidats

descripteurs se trouvent donc identifiés et situés dans la continuité du document. La matrice permet ainsi de :

-stocker le nombre d'occurrences des différents descripteurs de chaque UI. Nous attacherons une attention particulière aux descripteurs relatifs aux ENT.
-Compter les occurrences de chaque candidat descripteur;
-observer la distribution de ces occurrences dans le document pour évaluer leur portée dans les unités structurelles.

Cette phase permet d'effectuer une pondération en fonction de la nature du syntagme et une consolidation en fonction de son ''étendue''. La pondération considère par exemple, qu'une unité de nature titre ou extraite d'une surcharge textuelle de croquis, a une valeur informative plus importante qu'une unité apparaissant dans un fil du texte. La consolidation consiste à voir comment un terme est représenté non pas seulement dans une petite unité mais sur un ensemble d'éléments structurels. Il s'agit d'évaluer la portée de l'unité d'indexation dans un ou des blocs d'information donnés. Si un même syntagme se réfère à plusieurs unités structurelles, proches ou lointaines ; unités qui peuvent être textuelles et non textuelles, ceci est une indication sur les relations respectives de ces éléments structurels.

À terme, ces deux opérations doivent permettre, lors d'une requête, d'indiquer les UI ou les blocs d'information non seulement susceptibles de satisfaire cette requête, mais de classer ces requêtes suivant leur degré de pertinence.

Exploitation de la matrice pour produire l'index du corpus

La mise en évidence de la structure logique du document multimédia, ou des constituants linguistiques dans cette structure (analyse linguistique) à travers la matrice, permet non seulement la localisation précise de clés d'accès à l'information du document mais aussi des calculs sommaires de regroupement d'occurrences. L'organisation des résultats de cette démarche constitue l'étape de la construction de l'index du document hypermédia.

Par rapport à l'index traditionnel qui présente pour chaque descripteur retenu, les pages où il apparaît, l'index que nous proposons cherche à faire correspondre à chaque descripteur retenu, des parties pertinentes du document. La portée d'une référence est définie non pas comme une simple localisation, mais comme une zone réductible à une simple unité d'information ou étendue à n'importe quelle partie définie dans la structure logique du document.

Mise en œuvre de la méthode

La structure d'un document avec le développement de l'informatique, s'établit en fonction de formats divers : DOC, HTML/XML, voire " propriétaire ". Pour nos exemples illustratifs, nous avons choisi un premier document sur support papier, de format DOC, et un second sur le WEB de format HTML

Étape 1 : Mise en évidence de la structure logique du document

La structure logique d'un document peut être appréhendée à travers la définition type minimale de document ci-après :
DTD " minimale " pour un document structuré purement textuel

```
<!DOCTYPE Document [
<!ELEMENT Document     –-     (Titre, SectNiv1+)          >
<!ELEMENT Titre    --     (#PCDATA)                >
<!ELEMENT SectNiv1     --     PassageTxt                  >
<!ELEMENT PassageTxt   --     (#PCDATA)                   >
<!ELEMENT SectNiv1     --     (Titre, SectNiv2+)          >
<!ELEMENT SectNiv2     --     PassageTxt                  >
<!ELEMENT SectNiv2     --     (Titre, SectNiv3+)          >
...
]>
```

Extension de la méthode aux éléments non textuels

Les corpus à notre disposition, sont des documents de maintenance de matériel technique et informatique, dans lesquels les éléments non textuels ont une valeur informative confirmée. Ceci nous conduit à proposer une extension de cette méthode à ces éléments non textuels (ENT). Il s'agit non pas de faire de l'analyse d'image, mais d'un essai de représenter ces images par les textes les surchargeant ou ceux les explicitant. Le premier exemple ici, est pris dans un extrait du document MotorLegend.

Au niveau de la structuration logique du document : la question est de savoir si nous considérons l'élément non textuel comme une unité à part entière ou dépendante du texte. Nous avons opté pour la première alternative. Cela signifie simplement que nous choisissons d'appliquer à l'image une même méthode d'approche que le texte, à savoir la considérer comme pouvant être décomposée en parties. Ceci est possible même si un élément non textuel est partie intégrante d'un texte.

On définit pour représenter structurellement un document, trois nouveaux éléments non terminaux (ElNonText, Source et FragTxt) et un nouvel élément terminal (#SOURCE). (#SOURCE) permet de mémoriser la référence externe à un fichier informatique contenant un élément non textuel, comme une photographie, un schéma, ou un diagramme (Fichier .jpeg, .gif…). ElNonText (pour élément non textuel) qui joue dans la structure un rôle similaire à l'élément textuel non terminal PassageTxt, comme constituant d'une section de niveau quelconque. Les éléments PassageTxt et ElNonText ne se distinguent pas par leur rôle mais par leur nature.

L'élément non terminal Source n'a pas de nécessité absolue mais est introduit pour une meilleure lisibilité de la DTD. FragTxt est introduit pour recevoir chacune des bribes textuelles qui peuvent apparaître sur une image : surcharge, voire texte figurant sur un objet photographié. On reste ainsi dans le même schéma de principe de représentation du document.

DTD “ minimale ” pour un document structuré incluant des éléments non textuels

```
<!DOCTYPE Document [
<!ELEMENT Document     –-     (Titre, SectNiv1+)          >
<!ELEMENT Titre    --     (#PCDATA)              >
<!ELEMENT SectNiv1     --     PassageTxt          >
<!ELEMENT PassageTxt     --     (#PCDATA)          >
<!ELEMENT SectNiv1     --     ElNonText          >
<!ELEMENT ElNonText     --     ( Source, Titre?, FragTxt* )
      >
<!ELEMENT Source     --     (#SOURCE)          >
<!ELEMENT FragTxt     --     (#PCDATA)          >
<!ELEMENT SectNiv1     --     (Titre, SectNiv2+)          >
<!ELEMENT SectNiv2     --     PassageTxt          >
<!ELEMENT SectNiv2     --     ElNonText          >
<!ELEMENT SectNiv2     --     (Titre, SectNiv3+)          >
...
      ]
```

La structure de présentation HTML (HyperText Markup Language) du document électronique MOTORLEGEND, nous a permis de situer les éléments non textuels dans la linéarité du document par rapport aux fragments de texte. Nous avons ainsi repéré grâce à HTML, la première image comme la 5^{e} unité d’information (UI) de l’extrait, image intitulée *isetta.gif*

Une plage des 6 croquis suivants est considérée selon HTML, comme 3 UI images situées en 25, 32 et 36^{è} position dans la structure globale du document. Ces trois images nous les considérons comme structurées, dans une approche de découpage similaire au texte. Nous supposons que la surcharge textuelle des différents éléments non textuels, soumise à un logiciel de reconnaissance de caractère pourrait être extraite et constituer des candidats descripteurs intégrés parmi ceux du texte. Ceci est important, car nous avons fait le constat, certains candidats descripteurs du document n’apparaissaient que sur les ENT. Suivant la grammaire élaborée plus haut, nous pouvons transformer le document Motorlegend comme suit :

Représentation structurée de l’extrait du document Motorlegend : Section

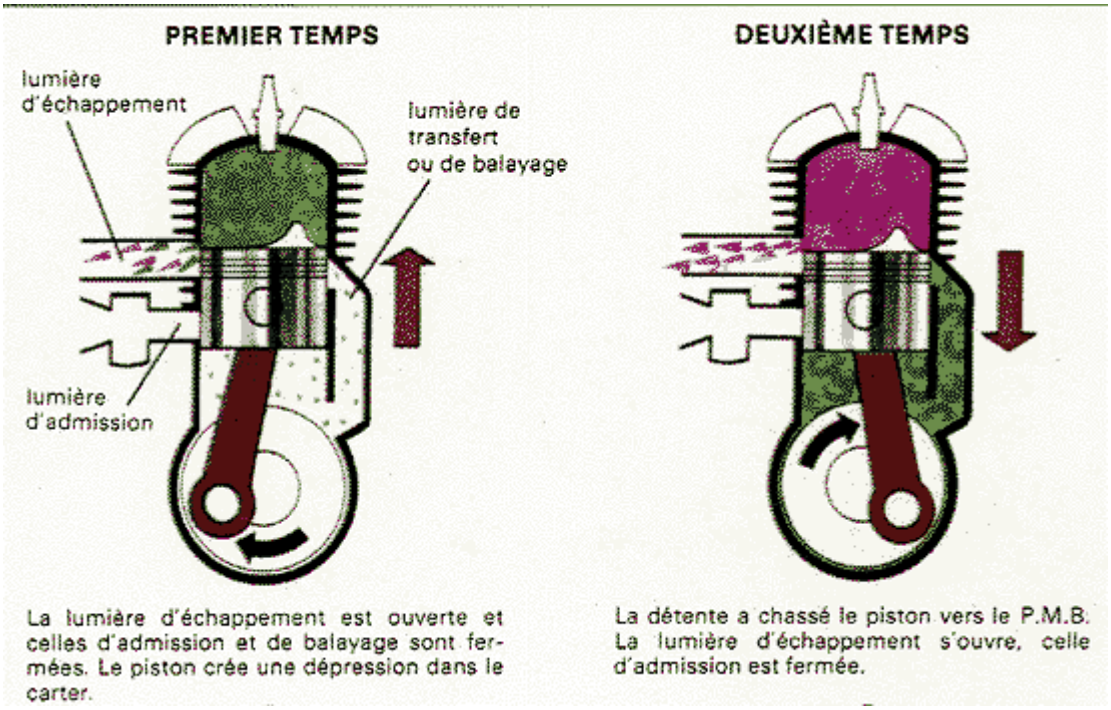
Cycle à deux temps

Le moteur à deux temps
On dit qu’un moteur alternatif à combustion interne…..vilebrequin……..
….Isetta.gif

Historique

……

Le cycle à deux temps


PREMIER TEMPS
DEUXIÈME TEMPS
lumière d'échappement
lumière de transfert ou de balayage
lumière d'admission
La lumière d'échappement est ouverte et celles d'admission et de balayage sont fermées. Le piston crée une dépression dans le carter.
La détente a chassé le piston vers le P.M.B. La lumière d'échappement s'ouvre, celle d'admission est fermée.

Document		{Le moteur à deux temps 1 et 2.html}
Titre		Le moteur à deux temps
Section		
	PassageTxt	On dit ... vilebrequin.
	ElNonTxt	{isetta.gif}
	PassageTxt	
	PassageTxt	
	PassageTxt	
	PassageTxt	
Section		
	Titre	Historique
	PassageTxt	
	PassageTxt	
	PassageTxt	
	PassageTxt	
	PassageTxt	
	PassageTxt	
	PassageTxt	
Section		
	Titre	Le cycle à deux temps
	PassageTxt	
	PassageTxt	
	PassageTxt	
	PassageTxt	
	ElNonTxtStruct	{1et2temps.gif}
	FragTxt	PREMIER TEMPS
	FragTxt	DEUXIEME TEMPS
	FragTxt	Lumière d'échappement

[…]

Étape 2 : Analyse linguistique

La structure du document établie, la seconde étape consiste effectuer l'analyse morphologique.

Analyse morphologique

Le principe est, à cette étape, d'associer à une forme textuelle, des caractéristiques permettant de repérer la base de cette forme, sous ses représentations diversifiées selon les contextes du discours. L'analyseur relève les variations de chacune des formes et propose, celle qui dans le contexte d'une unité d'information, en l'occurrence la phrase semble la plus pertinente. À titre d'exemple, nous donnons ci-après le résultat attendu de l'analyseur de l'équipe CRISTAL que nous avons manuellement simulé uniquement sur le titre du premier document d'essai.

Texte original	**Analyse morphologique**			
Vers	f*ers	P	vers	AAU LOC FAI
un	un	D	un	MAS SNG NNU FAI
modèle	modèle	F COM	modèle	MAS SNG NOMFOR
d'	de	P	de	APO FAI
indexation	indexation	F COM	indexation	FEM SNG NOMFOR
des	de	P	de	APO FAI
	les	D	les	GRN PLU DEF FAI
documents	documents	F COM	documents	MAS PLU NOMFOR
techniques	techniques	F	technique	GRN PLU ADJ FFN

Analyse « robuste » se substituant à l'analyse syntaxique

L'équipe CRISTAL a initié pour repérer les syntagmes devant servir de descripteur, différentes méthodes. L'une d'elle consiste à rechercher les types de procès du document. Virginia Bentès (1999) a ainsi élaboré une grammaire basée sur les schémas de structure des états et processus pour extraire les descripteurs. Céline Paganelli et Evelyne Mounier **(date)** propose ce même type de stratégie au niveau des unités d'interrogation au moment de la RI. Une seconde méthode part d'un analyseur syntaxique qui sectionne les phrases en propositions et donne leur nature.Pour cette méthode, une étape supplémentaire consiste à effectuer le choix des syntagmes nominaux. Les différents types de syntagmes nominaux proposés par l'analyseur sont très divers et posent un problème d'harmonisation.

Nous avons choisi de réaliser en une seule phase la détermination des descripteurs en proposant une analyse syntaxique que nous qualifions d'analyse robuste.

Principes

Le résultat de l'analyse morphologique se trouve en entrée de notre analyse linguistique robuste, qui en sortie doit nous donner la liste des candidats descripteurs avec leur localisation, c'est-à-dire leurs numéros d'identification.

Caractérisation des candidats descripteurs en sortie de l'analyse robuste

Les candidats consistent en un ou une séquence de lexèmes NOM.
Exemples :

Échappement
Tour/vilebrequin

Chaque lexème peut avoir une liste d'adjectifs.
Exemples :

Gaz (brûlé)
Moteur (alternatif) / combustion (interne)
Diesel (gros) / temps (deux, marin)

Automate pour l'analyse robuste

Nous utilisons pour réaliser cette extraction un automate d'états finis dont les constituants sont : les éléments de la chaîne d'entrée à analyser, les états, les transitions.

Les éléments en entrée

Ils sont constitués par une sous-détermination des catégories morpho-syntaxiques en seulement 5 éléments distincts :

- ♦ DET : les déterminants ;
- ♦ ADJ : les adjectifs ;
- ♦ NOM : les noms ;
- ♦ prep DE-A : les prépositions DE et A ;
- ♦ divers : Les autres catégories : verbes, pronoms, relatifs, conjonctions relatives, subordonnées, ponctuation.

Exemple :

On	*dit*	*qu'*	*un*	*moteur*	*alternatif*	*à*	*combustion*	*interne*
Divers	Divers	Divers	DET	NOM	ADJ	prep DE-A	NOM	ADJ

Les trois états :

- État A : on passe dans cet état entre les lexèmes d'un descripteur (et a fortiori entre les descripteurs) ;
- État B : pour le descripteur en cours, un déterminant et/ou un ou des adjectifs antéposés ont été extraits mais le NOM correspondant pas encore ;
- État C : le NOM courant du descripteur en cours a été extrait (des adjectifs postposés peuvent encore suivre).

La catégorie Divers ramène à l'état initial.

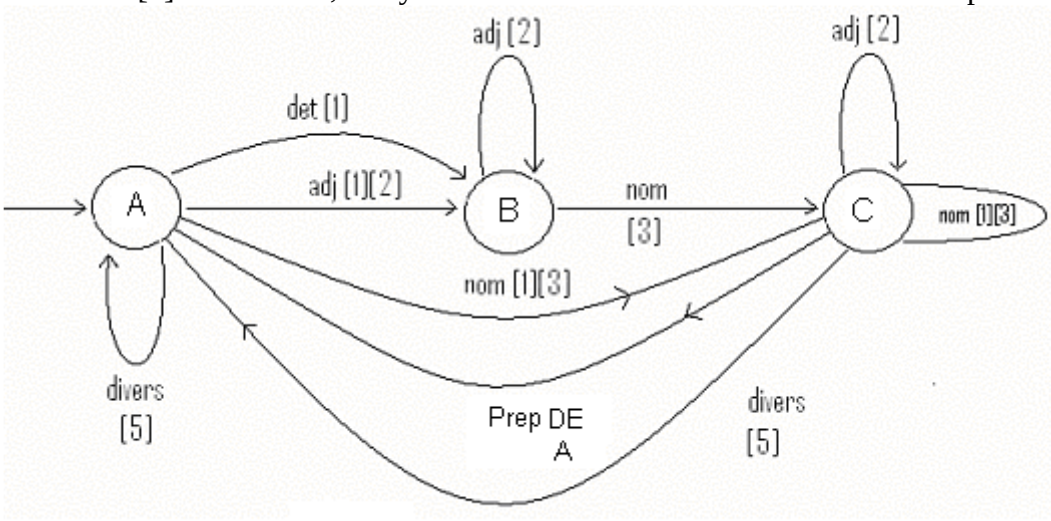
Les transitions

Le passage d'un état à l'autre est régi par l'état courant et l'élément en entrée (cf. schéma plus bas).

Les actions effectuées au passage des transitions

Elles sont les suivantes :

- [1] préparer une case pour un nom ;
- [2] ajouter un adjectif au nom courant (ou à venir) ;
- [3] stocker le nom dans la case courante du descripteur robuste ;
- [4] rien à faire ;
- [5] si case vide, il n'y a rien à faire sinon extraire candidat descripteur.



Exemple : entrée à analyser.

Le moteur à deux temps *O*n dit qu'un moteur alternatif à combustion interne fonctionne selon un ***cycle à deux temps*** lorsqu'…

Analyse robuste

État	**Entrée**	**Elément en entrée**	**Action**	**Descripteur en cours**	**Sortie de l'analyse robuste**	
					Candidat descripteur	**UI**
A	Le	Det	[1]			
B	moteur	Nom	[3]	Moteur		
C	à	Prep DE A	[néant]			

A	deux	Adj	[1] [2]	Moteur / ???(deux)		
B	temps	Nom	[3]	Moteur / Temps(deux)		
C	¶	Divers	[5]		Moteur / Temps(deux)	2
A	On	Divers	[5]			
A	dit	Divers	[5]			
A	que	Divers	[5]			
A	un	Det	[1]			
B	moteur	Nom	[3]	Moteur		
C	alternatif	Adj	[1] [2]	Moteur(alternatif)		
B	à	Prep DE A	[néant]			
A	combustion	Nom	[3]	Moteur(alternatif)/Combustion		
C	interne	Adj	[1] [2]	Moteur(alternatif) / Combustion(interne)		
B	fonctionne	Divers	[5]		Moteur(alternatif) / Combustion(interne)	4
A	selon	Divers	[5]			
A	un	Det	[1]			
B	cycle	Nom	[3]	Cycle		4
C	à	Prep DE A	[néant]			4
A	deux	Adj	[1] [2]	Cycle ??, (deux)		4
B	temps	Nom	[3]	Cycle (deux) /temps		
A	Lorsqu’	Divers	[5]		Cycle (deux) /temps	

Application à un document contenant des éléments non textuels

Au fur et à mesure du parcours des UI élémentaires composant le document, l’analyse robuste s’applique aux fragments textuels ou aux ENT. En ce cas, le nom du fichier source de l’ENT constitue un descripteur.
Exemple : UI 5 {isetta.gif} donnera : isetta.
En fonction de la nature du document électronique, soit on applique un logiciel de reconnaissance de caractères qui extrait la surcharge textuelle, soit on soumet directement cette surcharge ou la légende associée, à l’analyse robuste.

Étape 3 Post- traitement des candidats descripteurs

L’analyse robuste ramène donc les UI textuelles (et les parties textuelles des éléments non textuels) à la séquence de tous les candidats descripteurs qui y figurent. Cette représentation du contenu textuel peut être ensuite élaborée de plusieurs manières :

- plusieurs occurrences du même candidat (à des localisations différentes) doivent être regroupées ;

- à ce niveau, on peut opérer une élimination de certains mots jugés non informatifs* ;
- enfin, on doit pouvoir mettre en place l'accès aux informations à partir de candidats incomplets, capables d'obéir à un ordre alphabétique, les candidats descripteurs complexes posant des problèmes quant à ce classement.

Regroupement des occurrences

Plusieurs occurrences strictement identiques du même candidat (à des localisations différentes) doivent être regroupées.

Exemples de regroupements

No descripteur	**Descripteurs**	**UI de localisation**
…		
3	Cycle / temps (deux)	UI =4, 20, 21
…		
16	Admission	UI=6, 7, 9,21, 23, 31, 32
…		
18	Détente	UI=6,32
19	Échappement	UI=6, 7, 21, 32
20	Façon (différente)	UI=6
21	Cylindre	UI= 7, 21, 23, 24, 35
…		
24	Piston	UI= 7, 21, 23, 24, 34,35, 38

En revanche d'après nous, les diverses occurrences d'un nom dans des candidats descripteurs complexes non identiques mais approchants ne doivent pas être regroupées (par exemple *cycle* dans l'extrait précédent figure, à la fois dans *cycle / temps (deux)* et *cycle (complet) / travail*. D'une part, les variations flexionnelles ont déjà été unifiées grâce à l'analyse morphologique. Mais surtout, aller plus loin dans la fusion de candidats différents ferait perdre définitivement une information utile. En effet, la fusion consisterait à ne conserver que la partie commune de plusieurs candidats perdant ainsi la précision thématique de chaque descripteur.

Exemple : plusieurs candidats descripteurs contiennent le mot *moteur* :
- moteur / temps (deux) ;
- moteur (alternatif) / combustion (interne) ;
- moteur / essence ;
- fonctionnement/moteur.

Les unifier à *moteur* ferait disparaître toute la classification par type de moteur qui est sous-jacente ici. De plus dans l'étape d'organisation ultérieure, le lien entre chaque nom et les descripteurs complexes où il figure, sera établi.

Étape 4 : Organisation de l'index

Bien que la phase d'interrogation ne soit pas au centre de ce travail, la structure et l'organisation de l'index la conditionnent. Comment classer les descripteurs retenus ? L'ordre alphabétique appliqué à des descripteurs dont certains sont complexes, pose un problème : comment classer *combustion (interne*) ou *moteur / explosion* ?

Par ailleurs, il nous semble indispensable de mettre en place le moyen d'accéder aux descripteurs (en particulier aux descripteurs complexes) et à leurs localisations, à partir d'une partie seulement de ces descripteurs. En conséquence de ces 2 aspects liés, nous proposons l'organisation suivante :

- les descripteurs retenus seront numérotés de manière unique dans leur ordre de première apparition ;
- le nombre d'occurrences dans les différentes UI où ils figurent, sera mémorisé ;
- une liste alphabétique d'unitermes composant les différents descripteurs retenus, sera construite. Chaque uniterme pointera sur tous les descripteurs où il figure. Nous avons choisi de ne retenir comme uniterme, que les seuls noms et de ne pas faire figurer les adjectifs. En effet un adjectif détaché de tout nom ne peut permettre de caractériser un thème.

Exemple : extrait de la table des descripteurs retenus (Ordre de 1° occurrence)

Numéro descripteur	**Descripteurs**	**UI de localisation**
1	Moteur / temps (deux)	UI=2
2	Moteur (alternatif) / Combustion (interne)	UI=4
3	Cycle / temps (deux)	UI =4, 20, 21
4	Cycle (complet) / travail	UI=4
5	Course (deux) / piston	UI=4
6	Tour / vilebrequin	UI=4
6'	Isetta	UI 5
…	…	…
90	Début / allumage	UI 37
91	Temps	UI 37
92	Aspiration / mélange	UI 37
93	Phase (terminal)	UI 37
94	Cycle	UI 38

Table des unitermes (Ordre alphabétique)

Descripteur	**N° d'ordre hiérarchique**
1 2 temps	26
1 2 temps2	35
1 2 temps3	39'
Admission	16
Aspiration	94
Compresseur	36
Compression	63,86
Course	5

Cycle	3, 4, 94*
Cylindre	21
Début	92
Définition	43
Dépression	65,84
Détente	22
Diesel	9
Distinction	12
Échappement	19
Essence	8

Façon	20
Fin	22
Gaz	75,78, 89
Instant	62
Isetta	6’
Lumière	64
Mélange	86,87
Montée / piston	66
Moteur	1, 2, 8, 44
Point	25
Pompe	46
Pression	74
Principe	10

Réalisation	41
Rotation /	70
Soupape *	40
Temps	1,3,7,8,13,14,28 ,29,57,68,76,82, 83,91
Terme	7
Tour	6
Transfert	59
Travail	4
Utilisation	8
Vilebrequin	6,

Possibilités d’accès offertes par notre index

L’index ainsi conçu, est constitué par une table alphabétique d’unitermes. Chacun de ces unitermes pointe vers une ou plusieurs entrées de la table des descripteurs. A son tour chaque descripteur pointe sur une UI et fréquemment sur plusieurs où il apparaît alors comme élément unificateur de ces UI. Par le biais de tels descripteurs communs, des UI éventuellement éloignées se trouvent mises en relation.

Exemple de liens par les descripteurs : document Motorlegend analysé précédemment :le document Motorlegend se caractérise par une forte proportion de liens repérables à partir de descripteurs :

- dans un extrait de la section 19 : *le cycle à deux temps* , le terme *piston* apparaît dans les UI textuelles adjacentes 22, 23 (2 fois), 24(2 fois), puis dans la surcharge textuelle des UI non textuelles, 31, 32, 34, 35, 38. On constate ainsi l’existence de relations entre les UI 23-24 et ces UI non textuelles, repérables à partir du descripteur commun *piston.*
- De même, le descripteur *lumière de transfert* apparaît dans les UI textuelles adjacentes 23-24, puis en surcharge textuelle dans les UI non textuelles 30, 35, 37, 38.

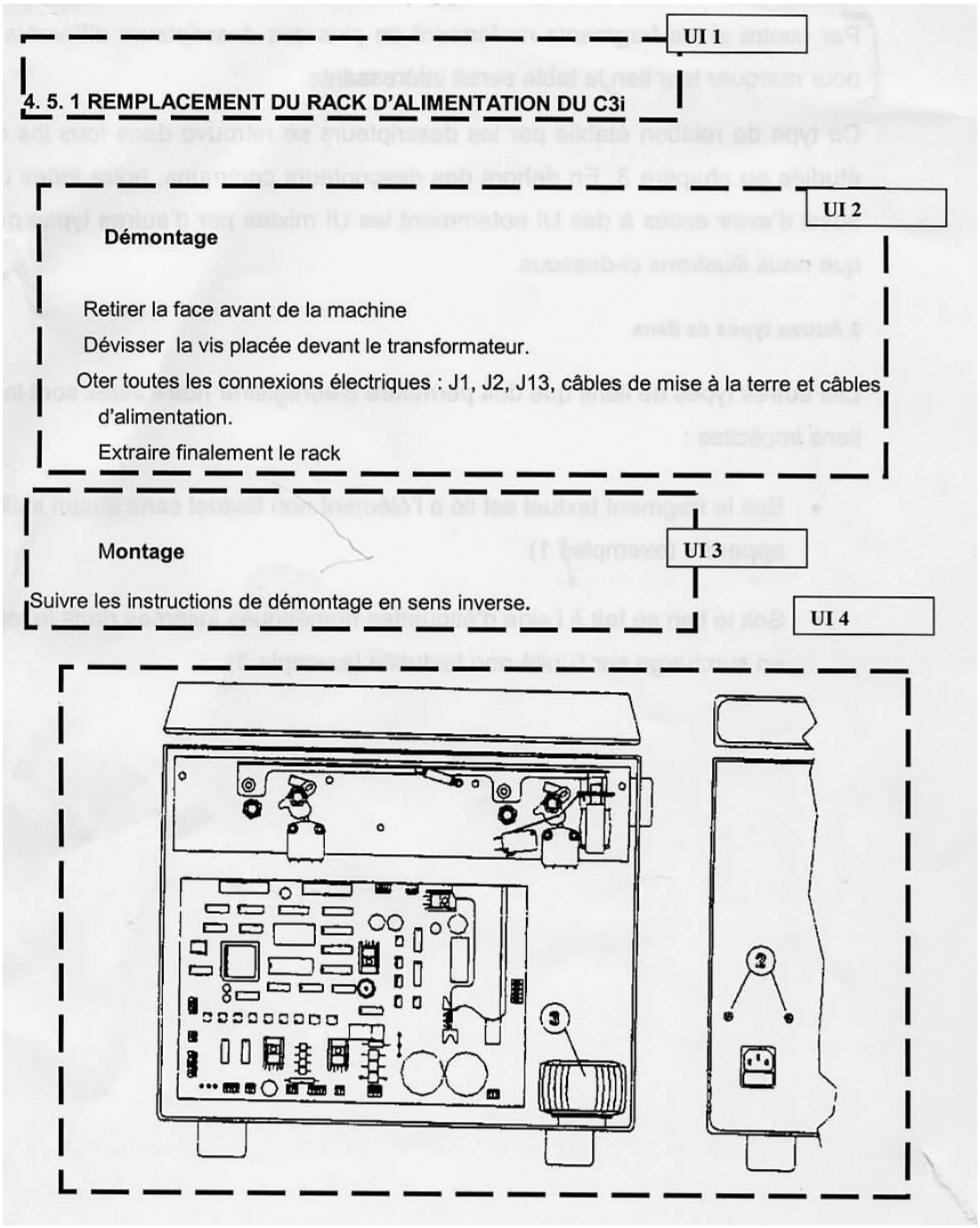
Les exemples ci-dessus concernent des relations entre UI proches et souvent adjacentes (cas largement majoritaire du document Motorlegend) avec des liens à l’intérieur d’une même unité structurelle : la section UI 19, titrée : *le cycle à deux temps*.
En conséquence, une requête qui renfermerait ces descripteurs communs pourra retourner les fragments textuels et non textuels les contenant, même si les fragments textuels ou non textuels se situent dans d’autres unités structurelles. En outre, une table qui enregistrerait ces relations s’avère non nécessaire. Par contre si les fragments renferment en plus d’autres descripteurs, alors pour marquer le lien entre ces UI, cette table a son importance. En dehors des descripteurs communs, notre index permet aussi d’avoir accès à des UI notamment les UI mixtes par d’autres moyens que nous illustrons ci-dessous.

Autres types de liens

Les autres types de liens que permet d’établir notre index concernent les liens implicites :

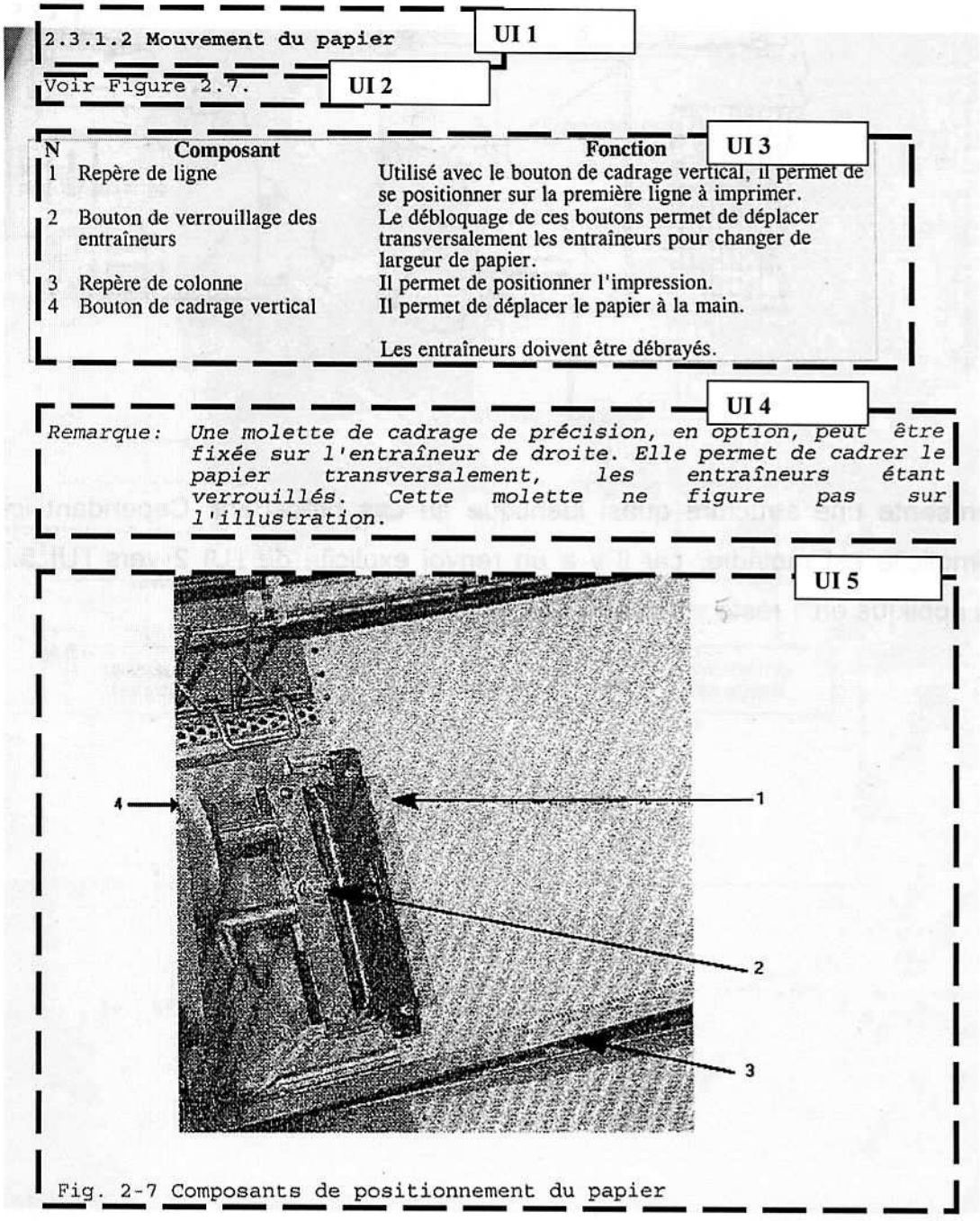
- Soit le fragment textuel est lié à l'élément non textuel sans aucun indice apparent

Exemple 1 Corpus Jouan


UI 1
4. 5. 1 REMPLACEMENT DU RACK D'ALIMENTATION DU C3i
UI 2
Démontage
Retirer la face avant de la machine
Dévisser la vis placée devant le transformateur.
Oter toutes les connexions électriques : J1, J2, J13, câbles de mise à la terre et câbles d'alimentation.
Extraire finalement le rack
Montage
UI 3
Suivre les instructions de démontage en sens inverse.
UI 4
2
3

Soit le lien se fait à l'aide d'étiquettes numériques insérées dans le texte et en surcharge sur l'unité non textuelle.

Exemple 2 Corpus Bull


2.3.1.2 Mouvement du papier
UI 1
Voir Figure 2.7.
UI 2
UI 3
UI 4
Remarque: Une molette de cadrage de précision, en option, peut être fixée sur l'entraîneur de droite. Elle permet de cadrer le papier transversalement, les entraîneurs étant verrouillés. Cette molette ne figure pas sur l'illustration.
UI 5
1
2
3
4
Fig. 2-7 Composants de positionnement du papier

N	Composant	Fonction
1	Repère de ligne	Utilisé avec le bouton de cadrage vertical, il permet de se positionner sur la première ligne à imprimer.
2	Bouton de verrouillage des entraîneurs	Le débloquage de ces boutons permet de déplacer transversalement les entraîneurs pour changer de largeur de papier.
3	Repère de colonne	Il permet de positionner l'impression.
4	Bouton de cadrage vertical	Il permet de déplacer le papier à la main. Les entraîneurs doivent être débrayés.

Dans le cas 1, comment la requête pourra lier les descripteurs des UI 1 à 3 à l'image ? (UI 4).

Dans le cas 2 comment lier *repère/ligne* ; *bouton/verrouillage/entraîneur*, etc., à la figure 2.7 qui n'a pas de surcharge, intuitivement reliée à la liste du tableau par les numéros (chiffres) des composants.

La structuration du document permet cette liaison. En effet, la phase de segmentation, dans une optique de consultation prévoit par un ensemble de règles des cas d'IU indissociables, constituant une unité documentaire (UD). Sous ce rapport, on lira, avec beaucoup d'intérêt Ouerfelli (2000).

Mémorisation des liens non hiérarchiques entre UI

Notre index permet l'accès aux unités d'information localisées dans la structure logique du document. Il sera possible d'afficher des unités mixtes en relation de séquence, c'est à dire des unités se suivant dans la linéarité du document soit grâce à des descripteurs, soit par les éléments structurels. Les relations transversales ne peuvent pas être repérées dans certains cas, pour les UI qui ne possèdent pas les mêmes descripteurs. Il nous semble indispensable donc le mémoriser dans une table de liens ces références croisées entre différentes UI : texte et image pour qu 'une requête puisse retourner des résultats constitués non seulement d'UI adjacentes ou repérables par des descripteurs.

Exemple

2.1 - PRESENTATION (se reporter aux figures du chapitre 6) **UI source 1**

Les centrifugeurs CR4-12, C4-12, CR3-12 et C3-12 sont des équipements fonctionnant sur les réseaux 120 ou 220 V.

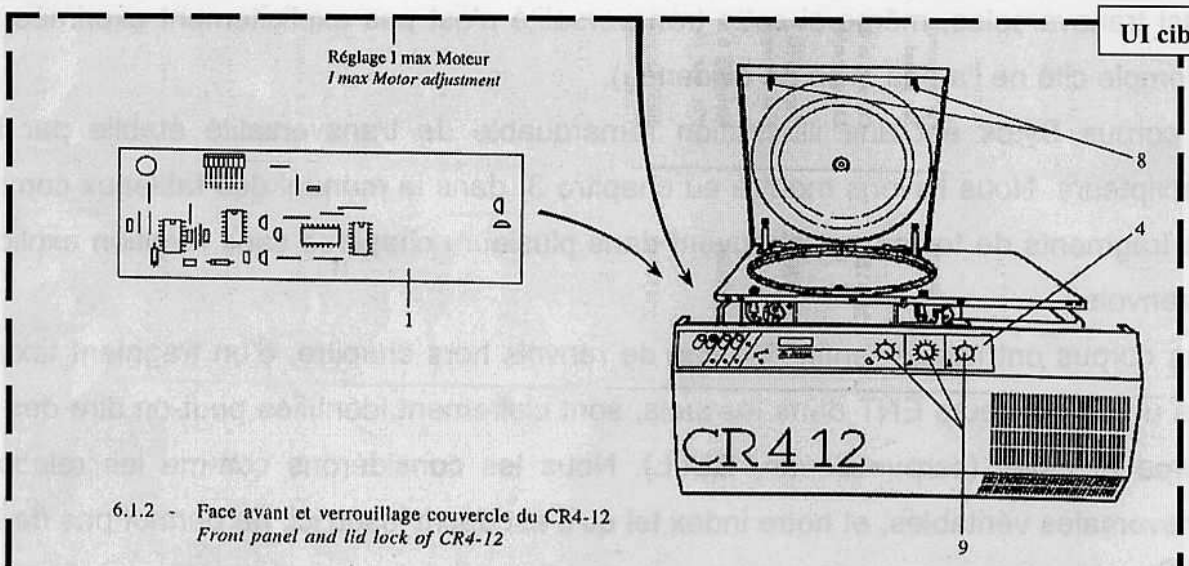
Les organes de contrôle-commande comprennent : **UI source 2**

- Un clavier souple (4, fig. 6.1.2 / 6.2.2 / 6.3.2 / 6.4.2),
- Un dispositif d'affichage numérique et des voyants de signalisation.

Le dispositif d'affichage numérique et les voyants de signalisation sont supportés par la carte afficheur (1, fig. 6.1.2 / 6.2.2 / 6.3.2 / 6.4.2).

Extrait de la page 6-6

UI cible 1



6.1.2 - Face avant et verrouillage couvercle du CR4-12
Front panel and lid lock of CR4-12

Extrait page 6-7

UI source

6.1.2 Face avant et verrouillage couvercle du CR4-12 (voir figure 6.1.2)

REPERE	REFERENCE	DESCRIPTION	Qté
1			
2			
3			
4	25370026	Clavier souple	1
5.....			

L'image de la figure 6.1.2, a une surcharge textuelle indirecte établie par les numéros fléchés (4,8,9) en rapport avec le tableau adjacent(extrait page 6-7). Les descripteurs qui en découlent ne permettent pas d'établir de lien avec les UI textuelles sources.

Représenter la liaison du chapitre 2.1 vers la figure du chapitre 6, dans une table doit permettre à une requête de retourner le fragment textuel titré en 2.1 et la figure et le tableau lié au chapitre 6. Ceci n'est pas possible à partir de la liste des unitermes établie précédemment. En effet, d'une part le titre du schéma de l'UI cible ne renferme pas le descripteur *clavier souple* contenu dans l'UI source : le paragraphe 2.1. D'autre part, la surcharge de cette UI cible, est faite de chiffres et le titre n'a aucun descripteur identique à celui de l'UI source. À partir du § 2.1, le tableau du chapitre 6 seul apparaîtra comme fragment autour du descripteur *clavier souple* et non la figure désignée par les renvois (6.1.2).

La représentation permettant la liaison de ces différents fragments ne sera possible qu'à partir d'une analyse linguistique des indices de surface liés aux renvois. On se reportera, par exemple, à la figure mise en évidence dans l'analyse de nos corpus.

Proposition de la table des références

Comment structurer la table des relations entre UI ? La liaison entre UI peut s'établir simplement en indiquant leur localisation dans la structure des UI source et cible. Cette localisation peut être représentée donc par une numérotation lorsque la fragmentation a été faite comme nous avons procédé sur le corpus Motorlegend, sinon on peut utiliser les références des UI.

UI source	**UI cible**
2.1	**6.6, 6.7**

Si l'on doit travailler sur une base de documents multiples, la distinction entre les UI des différents documents implique d'introduire un numéro de document dans l'identifiant des UI :

UI source	**UI cible**
1 - 2.1	**1 - 6.6, 1 - 6.7, 3 – 4.1**

Tout au long de cet article nous avons tenté d'effectuer l'indexation automatique d'un extrait de document ''technique '', suivant un modèle en quatre étapes. La phase 3 d'extraction des descripteurs a tenté de modéliser la pratique classique d'indexation consistant à extraire certains éléments du texte. Nous y avons ajouté une localisation plus fine des éléments extraits. Nous avons ainsi élaboré une grammaire minimale indépendante du domaine, pour segmenter le document, en nous appuyant sur la structure hiérarchique mais sans associer de signification particulière aux différents niveaux de cette structure. La

finalité de cette démarche est de construire un index permettant un accès à l'information de ce document dans ses parties les plus élémentaires.

Dans cette structure, ont été localisés les éléments non textuels (ENT), comme étant non subordonnés au texte (TXT), pour être soumis à une même méthode de segmentation que le texte. La structure interne propre des ENT, n'est pas systématiquement utilisée, seules les « bribes textuelles » incluses sont repérées. Par nécessité opératoire, nous avons ainsi séparé TXT et ENT. Nous pensons que lors de la phase d'interrogation, la possibilité d'agréger ces UI de nature différente s'impose et doit être réintroduite.

Une fois réalisée la localisation des UI, nous avons par une analyse linguistique en plusieurs phases, abouti à la représentation de ces UI en séquences de descripteurs liés aux syntagmes nominaux qui constituent les entrées de l'index. Nous avons successivement procédé à l'élimination des formes non désirées (verbes, déterminants, etc.) à partir d'un automate, puis établi les tables de descripteurs et d'unitermes affectés de leur(s) numéro(s) de localisation et pointant vers les descripteurs. En dernier lieu, nous avons enfin tenté de représenter les relations non hiérarchiques dans une dernière table.

La contribution que nous apportons aux travaux de CRISTAL, est d'offrir la possibilité d'inclure dans l'index des descripteurs relatifs aux ENT. Ces descripteurs permettent ainsi à des requêtes de retourner des éléments non textuels, non plus comme des UI fortuites apparaissant, parce que incluses dans des fragments de texte, mais comme des entités ayant même statut que ces fragments textuels.

Au terme de la description de notre modèle, qui clôt cette étude, notre conclusion portera sur deux aspects :

- les limites de notre outil principal d'analyse linguistique : l'automate. Nous parlerons en premier lieu des limites pour lesquelles nous envisageons des solutions. Nous identifierons ensuite celles pour lesquelles nous n'entrevoyons pas pour l'instant, de solution.
- Les problèmes relatifs à l'index.Tel que conçu, notre index est-il suffisant pour une RI efficace ou non dans le contexte de la documentation technique ? Quels problèmes posent une amélioration possible de cet outil ?

Limites de l'analyse robuste

Quelques propositions de solutions

Pour le passage de certains candidats descripteurs au statut de descripteurs, notre analyse linguistique subit un échec. Nous recensons ici ces cas posant problème et nous proposons des éléments de solution. Rappelons toutefois que nous considérons l'indexation comme un problème (non déterministe) qui peut se satisfaire d'une solution approchée (heuristique), au sens où une ou des solution(s) exacte(s) ne sont pas définies(s) de manière absolue.

Nous avons repéré dans l'extrait soumis à l'analyse, une majorité de mots reliés par les prépositions *de* et *à*, et avons essayé de les traiter en priorité. Concernant d'autres mots reliés par des prépositions (une infime minorité dans notre extrait) ou autres signes, constituant des mots composés, que pouvons nous dire ?

Les mots composés avec tiret (-)

Ce sont les mots composés classiques et dans nos corpus nous avons les exemples de mots composés tels : *carter-pompe*$^{(m)}$, ou *cylindre-pompe*$^{(m)}$. *moto-reducteur*$^{(j)}$.[4]. Soit, en sortie de l'analyse morphologique (AM), la forme composée a été étiquetée comme un seul élément que l'automate traitera comme tel (un descripteur et un uniterme) ; soit l'analyse morphologique a identifié 3 éléments consécutifs et produit deux formes. Dans notre cas, la version utilisée de l'analyseur morphologique (AM) de CRISTAL traite les mots comme un seul élément. Notre automate, devant le tiret (signe de ponctuation, catégorie Divers), analyse la forme comme deux descripteurs différents.

Il nous semble que l'un ou l'autre des deux traitements morphologiques distincts devraient pouvoir être appliqués suivant que le mot composé est attesté dans le lexique ou non. Un mot composé à tiret, lexicalisé serait ainsi traité correctement sans modification. En revanche, l'automate devrait être modifié pour traiter les mots composés à tiret mais non lexicalisés, de manière à produire le même descripteur que pour les compléments de nom avec *de* ou *à*.

Il faut donc distinguer le *mot-clé*$^{(*)}$, entrée lexicale (qui apparaît comme tel dans un dictionnaire) du *carter-pompe*$^{(m)}$, créé par l'auteur ou propre au domaine de la mécanique. Il y a la possibilité de modifier l'automate pour repérer la séquence de formes Nom-Tiret-Nom et identifier ainsi *carter-pompe*$^{(m)}$ à *carter/pompe*, comme s'il s'agissait de *carter de pompe*$^{(*)}$.

À propos de la nécessité de distinguer les mots composés à tiret lexicalisés de ceux qui ne le sont pas, se pose le problème d'interaction des phases de segmentation et d'analyse flexionnelle, phases que l'on sépare souvent par commodité.

Notons enfin que les mots composés sans tiret (–) (groupes nominaux lexicalisés comme *pomme de terre*, *carte bancaire*, etc.) sont normalement traités par le système proposé sans être explicitement identifiés comme mots composés.

Les noms propres lexicalisés et les mots inconnus

L'analyseur morphologique CRISTAL, identifie les noms propres à des mots inconnus. Ce choix justifié à l'origine en particulier par des raisons d'encombrement des dictionnaires, peut désormais difficilement être conservé.

On pourrait vouloir distinguer les noms propres « vrais » (principalement personnes et lieux), des mots créés « artificiellement » par l'auteur ou sa communauté (mots inconnus ou dans un lexique de spécialité). Mais la limite ne peut être fixée, en particulier, dans le domaine technique où on trouve des « noms propres d'objets » attestés dans le domaine.

En fait on doit/peut simplement distinguer les noms propres lexicalisés de ceux qui ne le sont pas. Dans chaque « catégorie de noms propres » (personnes, lieux, sociétés, marques, dispositifs, etc.), les 2 cas se présentent. Cependant, la couverture lexicale envisageable est plus ou moins grande suivant la catégorie :

- une majorité de prénoms peut être lexicalisée ;
- les noms de famille beaucoup moins.

[4] Les lettres en exposant désignent les corpus de notre étude,d'où sont tirés les exemples.

Les noms propres lexicalisés portant la catégorie nom ne posent pas de problème particulier lors de « l'analyse robuste ». Les mots inconnus (y compris les noms propres non lexicalisés, les sigles, etc.) peuvent aussi être inclus dans la catégorie Nom, sans problème comme ces deux exemples extraits du corpus JOUAN :

CR3i, IG150
EMKARATE RL

Remarques :

- les sigles posent un problème d'interaction entre segmentation de formes et segmentation de phrases lors de l'analyse syntaxique ;
- les noms (propres) de domaine ont des liens avec des thésaurus et « ontologies » ;
- les erreurs orthographiques contribuent à produire comme mots inconnus, des mots lexicalisés (noms communs comme propres).

La détection des erreurs orthographiques pour éviter de les prendre comme mots inconnus, est un problème sur lequel s'est penché Christel Froissart (2001) qui a élaboré une typologie ces erreurs. La question générale des noms propres doit être liée à une approche traitant les entités nommées dont nous parlons plus loin.

Les connecteurs et/ou

Les cas d'énumération (assez fréquents surtout avec la préposition *et*) s'analysent sans difficultés : Une fois ces limites atteintes, le ruban et l'encrier doivent ...[B] Dans de nombreux autres cas, notamment ceux dans lesquels interviennent des ellipses, l'analyse subit des échecs :

Lumière de transfert ou de balayage [M(ui30)]
Lumière d'admission et de transfert [M(ui30)]

Dans la réalité, ce candidat descripteur désigne t-il un seul objet ou deux ? L'analyse robuste donne les deux descripteurs :

Lumière/admission et *transfert. O*n attendrait *lumière/admission* et *lumière/transfert.*

Moteur élémentaire à trois lumières et à carter-pompe [M(ui21)]

Sur cet exemple, l'analyse robuste donne 3 descripteurs :

- moteur (élémentaire)/lumière(trois) ;
- carter ;
- pompe.

On s'attendrait plutôt à :

- moteur (élémentaire)/lumière (trois) ;
- moteu (élémentaire)/carter/pompe. ;

Ou même encore à : moteur(élémentaire)/lumière(trois)/carter/pompe.

[5] *Le monde motocycliste et en particulier cyclomotoriste* [M]

C'est un autre exemple dans lequel notre analyse échoue d'une manière un peu différente : le descripteur : *monde(motocycliste, cyclomotoriste)*, serait préférable à *monde(motocycliste)* que détecte notre automate alors que *cyclomotoriste* est perdu.

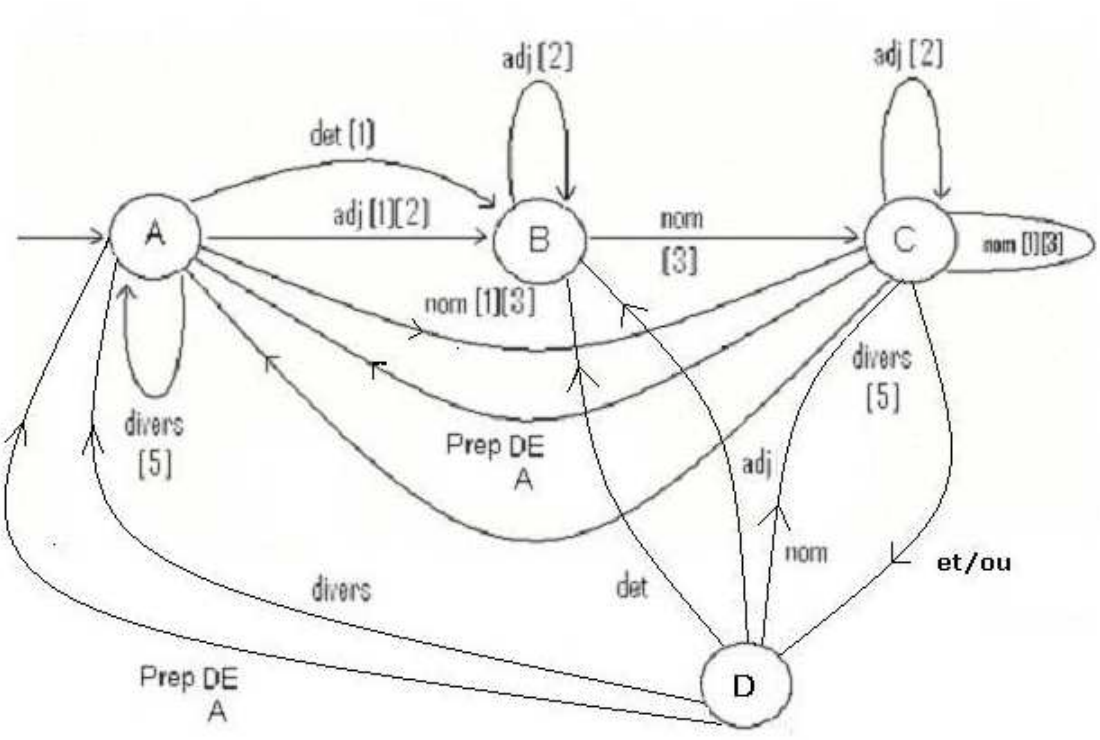
La modélisation linguistique de la coordination et son application en TAL est un problème difficile, en particulier parce que la coordination s'applique à des niveaux divers (propositions, groupes nominaux et verbaux, etc.). Ici encore, une solution partielle du problème linguistique général serait intéressante. D'une part, une solution limitée au contexte du groupe nominal est suffisante et d'autre part, en conservant l'idée d'analyse « robuste », une solution « approchée », efficace dans de nombreux cas, serait déjà productive.

Une telle solution partielle semble pouvoir être offerte par simple modification de l'automate d'analyse robuste dans le cas fréquemment attesté, de la succession immédiate de l'un ou l'autre des 2 connecteurs (Et, Ou)

Lumière de transfert ou de balayage *M(ui30)*
Lumière d'admission et de transfert *M(ui30)*

et de l'une ou l'autre des prépositions (De, À) :
Moteur (élémentaire à trois lumières ; Moteur élémentaire à carter-pompe

La détection par l'automate de l'un des 4 cas : *et à, et de, ou à, ou de*, est relativement simple. Elle consiste à dédoubler l'état A pour agir différemment lorsque l'on rencontre un des connecteurs :



Il est plus délicat de construire le traitement voulu ; il s'agit alors d'être en mesure de clore la construction du descripteur en cours (1) et d'initialiser un 2° descripteur à partir du descripteur précédent privé du dernier élément nominal (2).
Moteur (élémentaire)/lumière (trois) (1) ; moteur (élémentaire) (2) /carter/ pompe
Lumière/admission(1) lumière(2)/transfert.

Ceci pose la question de la portée de la coordination à l'intérieur même du groupe nominal. Cette difficulté peut être tranchée, sans grand risque, par le choix systématique de la portée maximale dans le groupe nominal pour le connecteur *et.* Un descripteur plus long est préférable à deux descripteurs au contraire du connecteur ou.. Ces principes émis, quelles nouvelles transitions rajouter alors aux transitions initiales si on envisage le traitement spécifique des connecteurs et et ou

- [1] préparer une case pour un nom ;
- [2] ajouter un adjectif au nom courant (ou à venir) ;
- [3] stocker le nom dans la case courante du descripteur robuste ;
- [4] rien à faire ;
- [5] si la case estvide, il n' y a rien à faire, sinon extraire le candidat descripteur et vider les cases ;
- [6]
    - si coord=ou alors extraire le candidat descripteur et vider seulement la dernière case (un nom + d'éventuels adjectifs) ,
    - si coord=et alors rien à faire (on continue la construction du même descripteur comme s'il n'y avait pas eu de *et*).

Quel état nouveau peut-on rencontrer ?

- État A : on passe dans cet état entre les lexèmes d'un descripteur (et a fortiori entre les descripteurs) ;
- Etat B : pour le descripteur en cours, un déterminant et/ou un ou des adjectifs antéposés ont été extraits mais le NOM correspondant pas encore ;
- état C : le NOM courant du descripteur en cours a été extrait (des adjectifs postposés peuvent encore suivre) ;
- état D L'un des connecteurs (*et* ou *ou*) a été détecté

La Lumière de transfert ou de balayage^{M}

Etat	**Entrée**	**Catégorie en entrée**	**Action**	**Descripteur en cours**	**Sortie de l'analyse robuste Candidat descripteur**
A	la	Det	[1]		
B	lumière	Nom	[3]	lumière	
C	de	Prep DE A	[4]	lumière	
A	transfert	Nom	[1][3]	Lumière / transfert	
C	ou	Conn ET OU	[4]	Lumière / transfert	
D	de	Prep DE A	[6]	Lumière	Lumière/transfert
A	balayage	Nom	[1][3]	Lumière / balayage	
C	<fin>	Divers	[5]		Lumière/balayage

La Lumière de transfert et de balayage*

Etat	**Entrée**	**Catégorie en entrée**	**Action**	**Descripteur en cours**	**Sortie de l'analyse robuste Candidat descripteur**
A	la	Det	[1]		
B	lumière	Nom	[3]	lumière	
C	de	Prep DE A	[4]	lumière	

A	transfert	Nom	[1][3]	Lumière / transfert	
C	et	Conn ET OU	[4]	Lumière / transfert	
D	de	Prep DE A	[6]	Lumière / transfert	
A	balayage	Nom	[1][3]	Lumière / transfert / balayage	
C	<fin>	Divers	[5]		Lumière / transfert / balayage

Le traitement proposé des deux connecteurs n'épuise pas la question et tout un éventail de difficultés subsistent:

Si la configuration matérielle a été modifiée ou qu'une version plus puissante a été installée...

Vous pouvez exécuter votre sauvegarde aussitôt que la modification ou la création et le test de la procédure de sauvegarde ont été terminés. [S]

Dans la mesure où notre automate ne prend en compte l'aspect sémantique à proprement parlé, la question relative aux connecteurs ne peut être tranchée. En effet, les valeurs sémantiques du ou, en français ont été étudiées par l'équipe CRISTAL à travers le mémoire de DEA de Claire Boudol.

L'étude met en évidence le fait que généralement l'exclusion est admise comme caractéristique du ou, alors que l'analyse du corpus traité montre la valeur additionnelle du ou comparable à la valeur additionnelle du mot et (Boudol 2002).

Un autre niveau de difficulté est celle relevée ci-dessous : exemple de la *Non concordance des paramètres matériels et des paramètres de l'unité de sauvegarde* [S]. On relève un cas attesté présentant une difficulté non résolue correctement par la modification de l'automate : *Non concordance des paramètres matériels et des paramètres de l'unité de sauvegarde* [S]
L'automate donne : concordance/paramètre(matériel)/paramètre/unité/sauvegarde.
Dans le cas, le sens même du mot *concordance* implique 2 compléments, ce qui est incompatible avec notre modèle de descripteur sous forme d'une séquence de nœuds nominaux. Traiter ce genre de cas nécessiterait une analyse sémantique non envisagée ici.

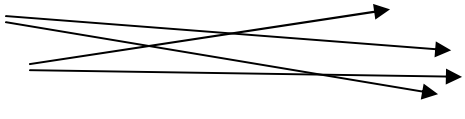
Autres problèmes soulevés par l'analyse robuste

Pour les problèmes soulevés ci-dessous, nous ne proposons pas de solutions au niveau de notre automate.

Les variantes de descripteurs

Dans le corpus Motorlegend, nous avons pour désigner l'objet *moteur à deux temps*, trois entrées différentes dans la table des descripteurs, entrées elles-mêmes pointées par les deux unitermes *moteur* et *temps* :

Uniterme	Descripteur	Venant de
moteur temps	Temps (deux) moteur/temps(deux) moteur (deux)/temps	*Le deux temps* *Le moteur à deux temps* *Le moteur deux temps*



Ceci présente deux inconvénients : d'une part les descripteurs sont inutilement multipliés, et d'autre part l'uniterme *moteur* ne permet pas d'accéder aux occurrences de *Temps (deux).*

Les reprises anaphoriques : Exemple dans le corpus Motorlegend

Ceux destinés à l'agriculture

Il s'agit ici de moteurs destinés à l'agriculture et même plus précisément de moteurs à deux temps. Au lieu du descripteur *moteur/agriculture* qui serait pertinent, l'analyse robuste va produire le descripteur simple *agriculture*. Dans l'index de ce corpus, *agriculture* se révèle un ‘’corps étranger’’. L'erreur est double : l'anaphore n'a pas pu être traitée et elle a engendré la présence d'un descripteur non pertinent dans l'index.

Signalons pour cet exemple précis, qu'il provient du corpus le ‘’moins technique’’ de nos documents, à visée encyclopédique. En documentation technique « pure », les normes de rédaction tendent à réduire l'apparition des reprises anaphoriques, sans toutefois les éliminer.

Le cas de la chaîne : *‘’leur développement’’* au lieu de : *développement de moteur**, va donner le descripteur simple *développement* et poser un problème de pertinence (moindre tout de même) comme dans le cas précédent.

Mots contextes

Le dernier exemple d'anaphore cité précédemment : *’leur développement’’* a une autre caractéristique révélée également dans l'expression : *possibilité de fonctionner* (qui à son tour pose le problème de verbalisation. Ces expressions permettent l'identification de descripteurs simples correspondant à des mots ‘’signaux’’ (Eymard 1998) ou des mots "contexte" (Bourigault et Charlet 1999). Ces descripteurs ne revêtent un sens que dans un contexte précis et en particulier ne désignent pas un "objet" auquel le discours réfèrerait. Leur maintien dans l'index n'est pas nécessaire.

Verbalisation / nominalisation

L'expression *possibilité de fonctionner* : est équivalent à *possibilité de fonctionnement*, et devrait donner le descripteur *Possibilité/fonctionnement*, ce qui n'est pas le cas. Cette question semble pouvoir être traitée en disposant dans le lexique de la forme nominale à laquelle certains infinitifs renverraient. Des descripteurs pourraient ainsi être construits par une analyse « robuste » adaptée. Mais une telle « solution » n'est pas satisfaisante. En effet, les descripteurs issus de l'analyse robuste sont censés représenter des « objets », ce

qui n'est pas le cas ici. La question doit pouvoir être abordée dans le cadre de l'analyse des catégories de procès (états, processus [Paganelli et Mounier **sans date**]).

Parenthèses et ponctuations en général

Nous avons constaté dans tous les corpus étudiés, que la parenthèse était fréquemment utilisée. Et contrairement à des corpus généraux dans lesquels le parenthésage intervient au niveau de propositions, il se trouve souvent au niveau de groupes nominaux dans nos documents analysés et joue des rôles variables :

- Précision sur la nature de l'élément avant la parenthèse

....on y retrouve les quatre phases classiques (admission, compression, détente, échappement), mais agencées…

… se situent entre la fin de la détente et le début de la compression (lorsque le piston est près du point mort bas).

- espèce de :

… n'exclut pas l'utilisation de soupapes (commandées ou automatiques).

- Marque simple de référence par les renvois :

…Pour le test qualité d'impression (voir chapitre 4…)

Les problèmes sous-jacents à ces exemples sont d'abord liés à l'extraction de descripteurs par l'automate puis à un tout autre niveau, relatifs à la construction de l'index.

L'extraction de descripteurs

Nous avons vu un exemple du document Motorlegend *: Soupapes (commandées ou automatiques),* nous perdons les qualificatifs rattachés à *soupapes* à cause de la parenthèse sous catégorisée Divers.

Dans l'exemple 1:

... puisqu'on y retrouve les quatre phases classiques (admission, compression, détente, échappement) $^{(m)}$

L'analyse ne permet d'avoir que des unitermes alors qu'on devrait avoir des descripteurs à deux ou trois termes : *phase (quatre,classique)/admission* ; *phase/admission* ; etc.

La virgule pose le même type de problème que les connecteurs.

Exemple

L'imprimante comporte des dispositifs de détection de bourrage papier, de fin de papier et de rupture de doigts porte-caractères$^{(B)}$

Analyse par les automates

Automate 1	Automate 2	Forme idéale
dispositif/détection/bourrage/ papier	dispositif/détection/bourrage/ papier	dispositif/détection/bourrage/
Fin /papier	Fin/papier	dispositif/détection/Fin papier
Rupture/doigt/porte/caractère	Fin/papier /Rupture/doigt/porte/caractère	dispositif/détection/rupture/doigt/porte/caractère

Remarque : ces cas restent marginaux dans nos corpus.

Les nombres et les dates

Nous n'avons pas proposé d'approche spécifique pour les nombres et les dates dans l'analyse de notre fragment Motorlegend. C'est évidemment un manque important. La question devra être abordée sur la base de travaux de recherche proposés dans la thématique dite des « entités nommées » (et en lien avec noms propres et mots inconnus). Là encore la difficulté tient à ce que la question a des répercussions à divers niveaux d'analyse linguistique (analyse graphique, syntaxique) et au niveau de la représentation des connaissances.

Dans le contexte d'une documentation technique de maintenance et d'utilisation, l'aspect historique qui pourrait faire envisager la question de l'extraction de date, est mineur. On pourrait inclure les dates dans la catégorie *Divers* de manière à ce qu'elles n'apparaissent pas dans l'index. Le corpus Motorlegend est, de ce point de vue, un cas limite comme nous l'avons déjà souligné. En revanche, au niveau des nombres (dates exclues), la question est à prendre en compte. En effet, nous avons des termes relevés dans le Corpus Motorlegend comme :

- *Phase de compression sur 180° de rotation du vilebrequin ;*
- *Fiat bicylindre 1500 $cm^{3,}$*
- *Un 500 cm^3 à deux temps*

Dans les cas, cités ci-dessus, le traitement des nombres doit être lié au traitement de désignation d'unités de mesure.

Les désignations spécifiques et formules de différentes disciplines

Exemples extraits du corpus JOUAN

Analyse de CO^2,
incubateur à CO^2
valeur mesurée par le module infra-rouge ($V= f(\%CO_2)^J$

Exemples extraits du corpus Motorlegend

La puissance en chevaux est donnée par la formule :

$$P = \frac{\pi \, n \, C^M}{30 \times 75}$$

L'identification de certaines de ces formes à des mots inconnus permet dans certains cas l'extraction de descripteurs corrects (*analyse/CO^2, incubateur/CO^2*). Les cas plus complexes tel la fraction précédente méritent une étude approfondie et systématique.

Conclusion

Nous avons essayé de concevoir un outil pour réaliser automatiquement la tâche d'indexation des documents techniques sur supports électroniques. La méthode se base sur la structure et l'analyse linguistique de ces documents, sans tenir compte du sens. Les problèmes qui se posent pour une utilisation optimale de cet outil ont trait justement à la prise en compte de cette sémantique que nous avons volontairement rejetée. Le corpus étudié est constitué de documents peu argumentatifs et cependant nous constatons, de

nombreux écueils liés à la langue, difficiles à surmonter par notre méthode. Que dire des corpus de documents non techniques ?

Le cas des noms propres et inconnus, des chiffres etc. ont été pris en compte et traités dans l'analyse des entitées nommées (named entity). Béatrice Daille e Emmanuel Morin font un ‘’bilan de récentes réalisations concernant la reconnaissance et la catégorisation des noms propres dans le cadre des Message Understanding Conference (MUC) (Daille et Morin 2000 : 621) initiées depuis les années 1980. Ainsi, les problèmes liés aux difficultés provoquées par l'emploi de ces expressions linguistiques particulières se trouvent regroupés sous la désignation d'entités nommées et font déjà l'objet de traitements à base statistique et linguistique.

De plus en plus, dans le cadre même de la documentation technique, on parle de représentation ou ‘’ingienierie des connaissances’’. Cette représentation finalement modélise, de façon plus fine et plus élaborée que ne le fait le thésaurus classique, les concepts d'un domaine donné, posant ainsi le problème de la fidélité au document représenté. Notre modèle privilégie cette fidélité au document.

Références bibliographiques

Balpe, Jean- Pierre et *al*. 1996, *Techniques avancées pour l'hypertexte*, Paris, Hermès.

Bentes Pinto, Virginia 1998, « La représentation des connaissances dans le contexte de la documentation technique : Proposition d'un modèle d'indexation », Grenoble,université Pierre Mendès- France, UFR [Thèse, Sciences de l'Information et de la Communication].

Boudol, Claire 2002, « Analyse de corpus, dans une perspective du traitement de la coordination en ou en génération automatique de textes », [Mémoire de DEA, Sciences du langage].

Bourigault, Didier et Charlet, Jean 1999, « Construction d'un index thématique de l'ingénierie des connaissances », *Plate-forme AFIA*, juin : 107-118.

Cotte, Dominique 2000, « Représentation des connaissances et convergence numérique : le défi de la complexité », *Document numérique*., vol.4, n°1-2 **: … ?**

Daille, Béatrice et Morin, Emmanuel 2000, «Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations », *Traitement Automatique des langues*, vol. 41, n°3 : 601-622.

Eymard, G. 1992, «Traitement documentaire des sommaires : des mots-clés à l'extraction des connaissances : Application à une documentation technique », Grenoble, Université Pierre Mendès- France [Thèse de doctorat, . Sciences de l'Info. et de la Communication].

Froissart, Christel 2001, «De la communication homme-machine à la recherche d'information dans la documentation technique : modélisation linguistique pour l'analyse automatique du français », [Mémoire de DEA Habilitation à diriger des recherches]

Lallich,-Boidin, Geneviève, Henneron, Gérard, Palermiti, Rosalba 1990, « Analyse du français : achgèvement et implantation de l'analyseur morpho-syntaxique », *Les Cahiers du CRISS*, n°16 **: …**

Ouerfelli, Tarek 2000, «Segmentation des documents techniques composites dans une perspective d'indexation : Vers la définition d'un modèle dans une optique d'automatisation ». Grenoble, UFR Sciences de la Communication et des Médias [Thèse de doctorat].

Paganelli, Céline, Mounier, Évelyne «Vers un système de consultation des documents volumineux par de utilisateurs experts :SYSRIT (Système de recherche d'informations techniques) », *Actes du Congrès d'Isko- France* : 195 - 225.